Incentives in the Public Sector: Evidence from a Government Agency

Simon Burgess[a], Carol Propper[b], Emma Tominey[c]

January 2016

Acknowledgements: We would like to thank two anonymous referees for their unusually detailed comments through the revisions. This work was funded by the Department for Work and Pensions (DWP), the Public Sector Productivity Panel, the Evidence-based Policy Fund and the Leverhulme Trust through CMPO. Tominey thanks the British Academy for funding. The views in the paper do not necessarily reflect those of these organisations. Thanks to individuals in the DWP for helping to secure the data for us, particularly Storm Janeway, Stavros Flouris and Phil Parramore. Thanks for comments to seminar participants at Bristol, the Public Economics Working Group Conference at Warwick, the IIES at Stockholm, University of Melbourne, HM Treasury, CPB in The Hague, Tinbergen Institute and Department for Work and Pensions.

Corresponding author: Emma Tominey. emma.tominey@york.ac.uk. Tel. +441904323781, Fax. +441904323759.

a *University of Bristol, CMPO, UK and CEPR, UK*
b University of Bristol, CMPO, UK, Imperial College London, UK and CEPR, UK
c University of York, UK, CMPO, UK and IZA.

Abstract

We study the impact of team-based performance pay in a major UK government agency, the public employment service. The scheme covered quantity and quality targets, measured with varying degrees of precision. We use unique data from the agency's performance management system and personnel records, linked to local labour market data. We show that on average the scheme had no significant effect but had a substantial positive effect in small teams, fitting an explanation combining free riding and peer monitoring. We also show that the impact was greater on better-measured quantity outcomes than quality outcomes. The scheme was very cost effective in small offices.

Keywords: Incentives, Public Sector, Teams, Performance, Personnel Economics

JEL classification : J33, J45, D23.

Governments employ a lot of people. The productivity of these workers, forming such a substantial fraction of the labour force (15% in the US, more than are employed in manufacturing at 13%) is therefore a major issue and many governments have an explicit agenda of improving the efficiency of public service delivery. One method which has received considerable attention is the use of explicit financial incentives. Early examples include Osborne and Gaebler (1993) "Reinventing Government", promoted by Vice-President Gore in the USA and the Job Partnership Agency Scheme in the USA in the 1980s (Barnow 2000, Heckman et al. 1996). More recently, there has been considerable interest in the use of performance related pay for teachers and public sector doctors and for outsourced firms providing services to the public sector (e.g. Muralidharan and Sundararaman 2011, Lavy 2009, Gravelle et al. 2010).[1]

Theorists have addressed the role of such incentives in the public sector, drawing attention to a set of features such as multiple tasks, multiple principals and missions. All of these suggest that even if there is no difference in the inputs of the production process between the private and public sector, there are some specific features related to outputs and to the way public sector agencies are structured which mean that incentives might be expected to have different consequences in public organisations (e.g. Dixit 2002, Prendergast 1999, Baker 2002, Francois 2000, Besley and Ghatak 2005). However, despite the interest (Burgess and Ratto 2003), the empirical evidence on the use of incentives in the public sector is still quite scant and a recent review concluded that there are relatively few estimates from which causal inferences can be made (Bloom and Van Reenen 2010).

This paper aims to fill this gap. In a search for greater public sector productivity, the government of the UK in the late 1990s set up a series of experiments into the use of financial incentives for lower level bureaucrats. As part of this programme they introduced an experiment in the use of team-based financial incentives in a large UK public agency. This agency, Jobcentre Plus, was one of the main government agencies dealing with the public and its role was to place the unemployed into jobs and administer welfare benefits. In contrast with many schemes in the private sector, the incentive scheme we analyse was exogenously imposed on the organisation as part of a wider government experiment with incentive pay (Makinson 2000). In addition, it was a team based performance pay scheme: workers were

[1] For example, in 2011 The UK government made payment- by-results a key part of its payment arrangements for private and not-for-profit firms for placing the long-term unemployed in work.

rewarded on the basis of team rather than individual production. If the team hit the targets set then all members of the team would receive the bonus payment.

We exploit this pilot to examine three key issues in the use of performance related pay in the public sector. First, what is the impact of an explicit financial incentive scheme on public sector workers? Dixit's (2002) review of theoretical contributions suggests that such incentives may be counter-productive. Evidence on the effectiveness of incentive schemes in the public sector is mixed. Kahn, Silva and Ziliak (2001) examine a reform to the Brazilian tax collection authority which paid financial incentives based on individual and team performance in detecting and fining tax evaders. Amounts involved were substantial, frequently providing bonuses over twice mean annual salary. Authors find a dramatic effect, with fine collections per inspection 75% higher than in the counter-factual. Lavy (2009) found that teacher incentives to improve pupil performance in maths and English in Israel significantly raised student outcomes. Baiker and Jacobson (2007) in a study of the police where participants were able to keep a proportion of the value of drug-related asset seizes found significant effect of the incentives, documenting an increase in heroin related drug offenses and even a rise in the price of heroin. Gertler and Vermeersch (2012) found improved child health outcomes from incentives to health workers in Rwanda to offer more and better quality of prenatal and postnatal care. But counter-examples exist – for example Mullen et al. (2010) found little effect of pay for performance on the quality of medical care.

Second, what is the impact of the team basis of the scheme? Economists have typically been skeptical of a team basis for obvious free-rider problems; such free rider effects have been found for example in Gaynor and Pauly (1990), Gaynor et al. (2004) and Bhattacherjee (2005). [2] On the other hand, Burgess et al. (2010) find that even in quite large teams, a team-based incentive scheme in the UK Customs and Excise raised the productivity of agency workers. Knez and Simester (2001) argue that peer monitoring outweighed free-riding effects in a scheme in Continental Airlines which had large teams and Hamilton, Nickerson and Owan (2003) conclude similarly for a garment factory in California. In an experimental setting, Carpenter (2007) and Abbink et al. (2006) find the relationship between group size and productivity to be dependent upon the design, for example the transparency of effort contributions. Ghatak et al. (1999) find that whilst in general smaller groups are favored for

[2] Holmström (1982) provides the formalisation.

joint liability programs in developing countries, how well group members know each other and interact are important. The incentive scheme we analyze was introduced across teams of very different structures, allowing us to quantify the effect of team size.

Third, how do workers respond to relative task measurement precision in an explicitly multi-tasking environment? Although the implications of multi-tasking for scheme design are a major part of the theoretical literature on incentives, there is very little empirical evidence on the importance of the precision with which targeted outcomes are measured (though see Gaynor and Pauly, 1990). The incentive scheme we study here incorporated five targets covering most of the tasks of the agency, though only four of these were measured. One was defined in terms of quantity of output and was measured with considerable precision as it was a weighted count of every client placed in a job. Three were defined in terms of quality and were measured with considerably less precision. We may therefore expect to find a greater impact on the quantity outcomes. On the other hand including targets for quality, albeit measured imprecisely, may prevent the decline in quality often associated with incentive schemes targeting quantity.[3]

The scheme was piloted in a small number of districts and we exploit plausibly exogenous assignment of treatment status for identification. The process by which some offices were given performance pay and others not is obviously key so we describe it in some detail. There were two unconnected events which generated a treated group. First, a small number of local offices were chosen to be a new kind of office (named Pathfinder offices). These were changed to offer a combined service of placing people into work and administering benefits, and were also given new IT systems to manage this. The Pathfinder offices were selected to be representative in terms of size and the level of urbanisation of their location, variables which we observe in the data. Clearly, the choice of offices to become Pathfinders was not random. Second, when the performance pay pilot was being set up, it was decided to introduce the incentive scheme in all offices in districts which had a Pathfinder office. Our analysis uses only non-Pathfinder offices. Our identification of a causal effect is therefore possible because for non-Pathfinder offices, whether they are selected for treatment or not is independent of their own characteristics. That is, their treatment status is exogenous. Ideally, we would want to run a difference-in-difference analysis but this is impossible. We only have

[3] See Paarsch and Shearer (2000) for an analysis of this issue.

data for the year that the scheme was in operation; consistent data for the year before simply does not exist as some district boundaries were re-drawn. We undertake propensity score matching to compare the incentivized offices with the most-alike non-incentivized offices.

We find no significant overall impact of the scheme. However, we do find significant heterogeneity of response that fits with free rider effects and the feasibility of peer monitoring in production. The impact of the incentive scheme was positive, substantial and significant in small offices and negligible in large offices. Thus while some mechanism such as peer monitoring does overcome the free-riding problem in small teams, it appears not to do so in large teams. Finally, whilst quantity increased, introducing multiple targets ensured there was no subsequent decline in the quality of service. Factors such as less precise measurement and poorer monitoring technology of quality relative to quantity meant that the scheme did not raise quality.

Section 1 describes the nature of the organisation and the incentive scheme and discusses the theoretical issues that arise. Section 2 introduces the data, and sets out our modelling framework and our identification strategy. Section 3 presents our estimation results. In section 4 we use these estimates to evaluate the scheme. Section 5 concludes.

1. The Organisation and the Incentive Scheme

1.1 The Nature of the Incentive Scheme

The initial drive for the introduction of financial incentives was political, originating in the White Paper "Modernising Government" (1999). This was followed up in the Makinson report (2000) for the Public Sector Productivity Panel, advocating incentive schemes for front line government workers. This study evaluates one of these schemes.[4] The pilot incentive scheme at Jobcentre Plus (JP) ran from April 2002 to March 2003. The main relevant features of the scheme are as follows.

[4] Burgess et al. (2010) evaluate another, implemented in the UK tax collection agency (Her Majesty's Custom and Excise).

1.1.1 Teams

Jobcentre Plus was organized in 11 regions and 90 districts[5]. Performance targets and rewards were assessed at the level of a district. A district contained a number of distinct offices, each dealing with the population in their local area. The official rationale for designing a team-based rather than an individual-based incentive scheme was to promote cooperation among workers, but discussions with the scheme designers revealed that another reason was that some of the output measures relating to the quality of the service provided by the agency are only available at the aggregate level of districts (see also below) [6].

All workers in the incentivized district got the bonus if the target was hit, including district managers and the district manager was responsible for achieving the target. Thus the team was defined on the basis of administrative structure rather than on the basis of a production function.[7] The teams were large. There were between 5 and 39 offices in the team and from 264 to 1535 people within a team.[8] 17 out of the 90 districts were incentivized. As noted in the introduction, these were the districts containing at least one new type of office – the 'Pathfinder' office.[9] These districts were designated as "Pathfinder districts" and contained both Pathfinder and non-Pathfinder offices. Non incentivized districts only contained non-Pathfinder offices.

1.1.2 Threshold incentive payment

In common with many schemes, the incentive scheme was a step function, based on a pre-set threshold level of performance. No workers were paid a bonus for performance below the threshold, the bonus was paid to all for hitting the threshold, and then there was no further increase in remuneration for further output. The bonus paid varied with the job grade and the number of targets hit, with a minimum of two targets required. There was an additional bonus for hitting all five targets. If all five targets were hit, a band A worker would earn an extra

[5] See Table W1 in the Web Appendix for a list of all districts and regions of Jobcentre Plus in 2002.
[6] It would have also been very hard to get union consent for the introduction of performance related pay based on individual output.
[7] There were few operational links between offices in a district and different offices were largely self-contained.
[8] Knez and Simester (2001) analyse the impact of incentives within big teams.
[9] These were offices that began to be introduced at the time when JP was launched in 2001, amalgamating the functions of two agencies: the Benefits Agency (BA), responsible for administering benefits to the unemployed, lone parents and others, and the Employment Service (ES), responsible for job placement. 56 new 'Pathfinder' offices were initially introduced to provide an integrated service, combining the work of the original, separate, benefits offices and employment offices. This process of change was slow, and most offices at the time of our study – there were 1464 in total – remained single service providers as ex-BA or ex-ES offices. More Pathfinder offices were created throughout the year of the pilot scheme.

£750, whereas a band G job would get £3,750 more[10]. This represented around 7.5 and 8.5 percent of average pay respectively. The threshold targets for the incentivized districts were set as percentage increases on the previous year's achievements.[11]

1.1.3 Multiple targets

One central issue in the design of incentive structures is the importance of multi-tasking. In particular, a trade-off between quantity produced and quality is often crucial. The scheme design recognized this and included targets for five different functions, which together measure quantity and quality. These were job placements (a measure of quantity), along with four measures of quality: customer service, employer service, "other business delivery functions", and reducing benefit calculation error and fraud. However, the specific activities involved, and the ease with which each target was measured, differed widely across the five targets. As workers will have to choose how to allocate their effort, the measurement of each target variable will affect the allocation of this effort.

The quantity measure, job placements (called job entries in the agency) were measured as weighted numbers of clients who were found work by the office. The weight per placement varied with the priority of the clients and reflected government targets (see Table W3 in the Web Appendix for details). Our main quantity output measure is job entry point productivity, defined as a simple ratio of total job entry points at office level divided by the number of frontline staff at office level.

A second measure, the quality of service to job seekers (denoted JSQ, also referred to as "customer service") captured aspects of quality - speed, accuracy, pro-activity of service, and the nature of the office environment[12]. It was measured by independent analysis of questionnaires to employers and 'mystery shopping' techniques.[13] The "employer quality"

[10] See Table W2 in the Web Appendix for more details on the actual bonus payments made to incentivized districts.

[11] All districts have clear targets set for all functions, but in the control districts these were not incentivized. The terminology of JP describes these base goals as targets and the higher levels as 'stretch'. In this paper we keep to the standard economics terminology and describe the higher levels of output required to win the bonus as the targets. See Table W7 for details of the incentive targets on each output measure.

[12] See Table W4 in the Web Appendix for more details.

[13] This consists of a quarterly programme, where the assessors used a variety of techniques to measure the elements of the target. In particular, they went into offices and acted out the role of a customer, checking the environment in which services were delivered and telephoned offices to see how quickly and effectively phone calls were answered.

target (EMQ) was a measure of whether and how quickly vacancies were filled[14]. This was measured (again independently) by a survey of employers. The "business delivery" target (BDT) covered a wide range of other functions, and appeared to be an attempt to measure everything else that the offices did[15]. It was measured by checking samples of cases. The final target, the "monetary value of fraud and error", focused on two particular benefits – Income Support and Jobseeker's Allowance. This was measured by specialist teams visiting each district and examining samples of cases. However, the measurement and tracking of this particular target was obscure, all 17 Pathfinder districts were treated as a single virtual region and reporting of progress on the target was well after the end of the pilot. Consequently the target provided no scope for policy evaluation within the period covered by the data and we ignore it here.

1.1.4 Hierarchy: measurement, reward and production.

A final relevant characteristic of the scheme was that targets were measured at different levels of the agency hierarchy. Job entries were measured monthly at office level. The three quality measures we examine, JSQ, EMQ and BDT, were measured quarterly at district level.

1.2 Theoretical Issues

While incentive schemes can impact on the selection of workers into organisations (for example, Lazear 2001, Dixit 2002, Besley and Ghatak 2003, 2005, Bandiera et al. 2011), the timescale of the pilot and the relatively low staff turnover we observe in the data suggest that, in this context, the main effects will come through changes in the behaviour of incumbent workers.

1.2.1 Structure and size of teams

An important characteristic of the scheme was the structure of the teams. These were defined at the level of a district and were made up of a number of offices with no *operational* link with each other. In our context, a classic Holmström (1982) team would be at the office level where workers depend on each other to produce output. But teams were defined by administrative boundaries (the districts), chosen as the units of measurement for targets and performance. This created interdependencies among the offices in a district. The expected

[14] See Table W5 in the Web Appendix for more details.
[15] See Table W6 in the Web Appendix for more details.

reward for effort in an incentivized office depended on performance at district level and this was determined by the output of all offices belonging to the same district. But production occurred at office level. Hence the structure of the scheme resulted in a two-level team: "natural" teams (offices) within reward teams (districts). At the level of an office, the fact that individual contributions to office output were not separately observable (as only a measure of the office output was available) creates an externality, similar to that of Holmström (1982) when output is fully shared among team members whose contributions are not separately observable and consequently the number of team members becomes crucial for the delivery of optimal incentives. In large teams the free-riding problem is more complicated to tackle with the use of group penalties/bonuses alone. Monitoring performance is also required. Peer pressure within an office - where colleagues are able to observe each other – could alleviate free rider problems but is more likely to be effective in small offices.[16]

In our context we have possible free-riding within an office *and* within a district[17]. It is not possible to identify the latter because we drop the PF offices for identification and in larger districts there tended to be more PF offices; thus this mechanical effect is confounded with any potential free-riding effect. Consequently our analysis of team size effects focuses on the office size effect (staff per office)[18]. We expect that offices with relatively fewer staff should perform better, as the free-riding issue is easier to tackle and peer pressure may be stronger.

1.2.2 Multi-tasking and the Measurement Technology

The main quantity target, job entry productivity, was measured most precisely and directly from the management information database, at office level, monthly. By contrast, the quality of service to job-seekers and employers were measured through a sample survey and a sample of cases, only at district level and only on a quarterly basis and contained an element of subjectivity. This greater level of aggregation over both time and space and increased uncertainty from the subjective element of assessment gives a noisier measure of how a

[16] Kandel and Lazear (1992) show that peer pressure can offset free-riding tendencies, but the strength of this peer pressure varies with unit size, with more effective monitoring in small units. Knez and Simester (2001) find evidence that free-riding can be reduced in large teams through team design.

[17] Ratto et al. (2010) provide a theoretical analysis of the different effects of size of office in the context of sub-teams operating within a larger team, where the reward is at the larger team level.

[18] We control for district size (number of offices per district) but it cannot be interpreted in terms of free-riding or peer monitoring.

worker's effort maps into output on these tasks. The enforcement of effort levels is even more difficult for the tasks measured at district level. When performance outcomes are low the district manager does not know which office is under-performing, making the coordination and monitoring more difficult and therefore free-riding across offices harder to tackle.

We can illustrate the optimal response of an employee given the reward structure and measurement technology in a simple framework[19]. Suppose an employee *i* produces output $x(a_i)$ depending on her effort level *a.* The output of each employee is sampled with probability *p*, and a performance measure is produced by averaging across *N* individuals. For simplicity assume that this aggregate indicator is measured with noise, denoted by ϵ. The target level *t* is hit when $(\sum_i px(a_i)/N) + \epsilon \geq t$ and the probability of achieving the target is therefore $F\left((\sum_i px(a_i)/N) - t\right)$. The marginal effect of worker *i's* effort on the probability of achieving the target is $f\left((\sum_i px(a_i)/N) - t\right).p/N$ where f is the density function of ϵ. Given that the rewards for hitting each target were the same and assuming no substantial differences in effort costs across targets, this allows us to predict how workers would have optimally focused their effort. We would expect more effort on the quantity target than the quality targets because: every worker's effort necessarily counted in the quantity target (p = 1) but only a sample was taken for quality ($p < 1$); the degree of aggregation of the performance information is higher in quality (district) than quantity (office); and the noise to signal ratio is lower in quantity measurement (high f in the neighbourhood of the target). However as workers had to hit at least two targets to receive any bonus, they could not fully neglect quality.

2. Data, Identification and Empirical Model

2.1 Data

We use data from JP's management information system and personnel database. These data were available only for the period of operation of the incentive scheme (April 2002-March 2003).[20] Management information recorded performance against the five targets. As noted

[19] We thank an anonymous referee for suggesting this way of illustrating the point.

[20] It would obviously be very desirable to have data before the scheme was implemented to allow a difference-in-difference technique. Unfortunately this did not exist: the district boundaries that defined the scheme were re-drawn in 2002.

above, Job entry productivity (JEP) achieved for each office on a monthly basis was the measure of quantity and the three quality outcomes (JSQ, EMQ and BDT) were reported for each district on a quarterly basis. A description of the data is provided in Table A1. It shows wide variation in quantity (JEP) across offices and time with the standard deviation of a similar magnitude to the mean but much less variation (and fewer observations) for the sample-based measures of quality relative to their mean.

The main input is people. We obtained, from personnel records, the number of staff in each grade for each office per month. The numbers in different grades appeared in more-or-less fixed proportions. For example, there was about one Executive Officer (EO) to two Administrative Officers (AO). Consequently, including numbers of each grade in the analysis leads to severe multicollinearity. We therefore defined a measure of front-line staff which was the office total of all workers in EO and AO grades.[21] For office level analysis we take the office mean of frontline staff across time.

We merged unemployment and vacancy data from the local labour market as a control for the difficulty of placing individuals in the labour market. Using the postcode (zip code) for each JP office, we located each office in a Travel To Work Area (TTWA).[22] We then extracted claimant inflow and vacancy inflow data for each TTWA and for each month.[23] We cannot take the unemployment and vacancy stocks as exogenous as they are influenced by the outflow rate, our dependent variable. So we use the inflow, both of unemployed claimants and of vacancies, and take the latter divided by the former. The state of the labour market plays two roles – first it provides the 'raw material' necessary for the office to produce job entries and second, it proxies labour market tightness and hence the ease of placing claimants in jobs. The office level mean labour market takes the average for each office over time.

Clearly the quality of the workers employed in the offices is an important consideration and there is no reason to expect it to be constant across the country. Traditionally, public sector jobs pay less than private sector jobs but variation in the differences between public and private sector wages across the country will feed into quality variation. To adjust for quality

[21] We have no information on the state of the capital (principally computing and communications equipment) in offices.
[22] These are largely self-contained local labour markets, defined by 75% of those living there also working there, and 75% of those working there also living there. There are some 400 covering Britain.
[23] National Online Manpower Information Service, http://www.nomisweb.co.uk/ .

we merge data on the local private-public sector wage differential as a proxy for the different quality of staff (see Nickell and Quintini 2002, Propper and Van Reenen 2010). From the Labour Force Survey Small Areas dataset we constructed the wage gap between the private sector and the public sector for each local authority using the difference in the hourly wage of full-time workers in the private sector and the public sector, measured in GB pounds. This was matched to the office postcode.

We know which offices were Pathfinder (PF) offices. They had to merge and deal with two separate functions which had previously been undertaken in separate offices. JP estimated that Pathfinder offices took at least five months to adjust, and even beyond the adjustment period, Pathfinder offices fulfilled more roles than a typical office. Consequently we would expect their productivity during the life of the pilot to be lower.

Figure 1 shows the distribution of the annual job entry productivity across different office and district types. Comparing offices in non-incentivized districts with non-PF offices in incentivized districts is a like-with-like comparison and the distributions are fairly similar. PF offices, on the other hand, clearly have lower mean job entry figures. As noted above, we exclude PF offices from our main analysis[24]. Figure 2 presents the distribution of the quality measures and shows higher attainment on the business delivery target, followed by employer quality and finally job seeker service quality.

2.2 Identification Strategy

The pilot incentive scheme was introduced in all offices in the 17 districts with Pathfinder (PF) offices, leaving all offices in the other 73 districts as controls. The choice of which offices were to become PF offices was made by Field Directors and their District Managers on the basis that their management would be able to cope well with the demands of the new structure. PF offices were located across the regions, and the selected offices were to reflect a "cross-section of different communities and customer bases, i.e. from large inner-city offices

[24] We do include another kind of office that during the scheme became designated as pathfinder offices (called Jobcentre Plus Offices). These became Pathfinders late in the time period of the pilot - 82% of these were not designated as Pathfinder until the final two months of the scheme – and given this we expect them to have had little disruption during the scheme. We test the robustness of this choice.

to those in smaller towns, suburbs and rural areas."[25] Clearly PF status is likely to be correlated with other inputs and outcomes.

Identification of a causal effect is possible because the treatment status of non-PF offices is unrelated to their own characteristics. Among the set of all non-PF offices across the country, some were given the incentive scheme and some were not, on a basis that was exogenous to their own inputs and productivity. PF offices themselves are clearly not randomly selected and they are omitted from the analysis.

One issue that might threaten identification would be the presence of important production spill-overs between offices: if the presence of a PF office affected the performance of other non-PF offices within the district, this would introduce bias. For example, a PF office may have absorbed clients from neighboring offices, or absorbed resources from local offices. In fact, the former is unlikely as roll out was quite disruptive and the latter also unlikely as the intention was for all offices to become Pathfinder offices. So as Pathfinder status was temporary there was little incentive to differentially shift clients or resources towards or away from other offices in the Pathfinder districts.

2.3 Matching offices

Table 1 shows that incentivized districts are larger (425 compared to 268 on average), both with more staff per office (32 compared to 27), and more offices (15 compared to 11). incentivized This is driven by the presence of the PF office(s); once these are dropped, the patterns are much more similar. They appear to face very similar labour market conditions. In panel b) of Table 1, incentivized offices have slightly more staff (32 compared to 27) when including PF offices but again, dropping these offices brings the mean office size of incentivized offices to 25. Again, labour market tightness is similar across treatment status. Comparing the variable for the private-public sector pay gap across treated and non-treated groups, we see only small differences across office and districts with incentivized districts excluding PF offices having slightly higher relative quality in the public sector but at office level slightly lower.

To further ensure that our comparison of incentivized offices with non-incentivized offices compares like-with-like we use propensity score matching to select our control offices. Even

[25] Private communication.

though districts are the basis for assignment into the treated category, we compute propensity scores at office level because offices are the unit of analysis. We include all non-PF offices in incentivized districts and all offices in non-incentivized districts, giving 912 offices. We estimate the conditional probability of assignment to incentivisation status based on a set of observable variables. These variables might influence choice of pilot areas and/or the outcome variables. We employ a nonparametric regression method with kernel weights proportional to an Epanechnikov kernel and bootstrap to calculate the standard errors using 100 replications with replacement.

To justify that our identification strategy compares treated and control offices which are similar in traits, other than assignment to the incentive scheme, the propensity score estimate from the above probit is shown in Table A2. Our identification strategy requires dropping PF offices from analysis. These offices are relatively large and therefore the consequence will be to lower the average office size and district size (through excluding offices) in treated teams and that is what we see in Table A2. The table indicates that it is important to ensure that each treated office has a comparable control through propensity score matching. Note that any difference in the private-public wage gap noted in Table 1 is insignificant in Table A2. Table 2 shows the balance of variables used for matching, across the treatment and control offices. Before matching, treatment and control means differ significantly in very few variables - the propensity score, the labour market variance and the number of offices per district. However, once we have implemented the matching technique, none of these differences are significant and indeed for almost all variables there is no significant difference between treatment and control. We therefore select the matching sample of 841 offices which were on common support for analysis.[2627] The evidence suggests that using this sample, our identification strategy is valid.

2.4 Empirical Specification

We aim to answer three questions. First, is the productivity of public sector workers influenced by financial incentives? Second, does free-riding matter in a team-based incentive scheme? Third, does targeting quantity and quality prevent a decline in quality despite differential measurement precision and poorer monitoring technology?

[26] See Heckman, Ichimura and Todd (1997)
[27] Table A3 details the estimation results from the propensity score matching.

The outcomes we focus on are log job entry productivity as the quantity measure and the quality of service to job seekers (denoted JSQ), the quality of service to firms (denoted EMQ) and the business delivery target (denoted BDT) as the three quality measures.

We estimate the quantity measures at office level

$$y_{od} = \alpha + \gamma IS_d + \beta X_{od} + \nu_{od} \quad (1)$$

where y is log total job entry productivity in office o in district d, X is a set of covariates and ν is random noise. γ denotes the effect of incentivisation status (*IS*), our parameter of interest. To test for the presence of free-riding within teams of the incentive scheme we interact *IS* with the number of workers within an office. We allow the treatment effect to vary flexibly across office size, by interacting *IS* with bins for office size, which compare the bottom and top quartile of office size (12 and 37 respectively) to the middle half of the data. We include an additional control for the interaction of *IS* with the number of offices per district as large incentivized districts tended to include a larger number of PF offices, which were excluded from analysis. Office level regressions cluster at the district level, given the importance of the district manager to coordinate the activity of the offices in their district. We also include regional fixed effects in all regressions.

We estimate the quality measures at district level:

$$y_d = \partial + \lambda IS_d + \eta Z_d + u_d \quad (2)$$

y={JSQ, EMQ, BDT}, λ denotes the effect of incentivisation and u the error term. To test for free-riding at the district level we interact IS with the number district staff. The controls *(Z)* are aggregated to a district level. We control for the size of the districts through the total number of staff in the district.

Our identification assumption for the office level analysis is:

$$E(y_{od}|IS_d=1,PF_{od}=0,X_{od})-E(y_{od}|IS_d=0,X_{od})=\gamma; \qquad E(\nu_{od}|IS_d,X_{od},PF_{od}=0)=0$$

and is the same for district level analysis, with the office subscript omitted.

3. Results

We present results first for the quantity variable, testing to see whether the incentive scheme had any effect on productivity and for evidence of free-riding. Next the effect of the scheme

on quality measures is assessed, to examine whether the effect of the scheme changes across outcomes measured with different precision.

3.1 Quantity – Job Entry Productivity

Table 3 presents the results for the log annual job entry productivity. This is for the sample of offices producing job entries, PF offices are excluded and we include only offices on common support in the matching. Standard errors are clustered at district level.

We present a number of different specifications for the effects of the incentive scheme along with the office characteristics. We start with the effect of basic office characteristics in column 1. Neither District Offices (which have central administrative functions) nor "Jobcentre Plus" offices (which became PF offices at the end of the year) have different job entry productivity from other offices. The main contextual variables have the expected signs but are insignificant: a high private-public wage gap reduces productivity, and a strong local labour market raises job entry productivity.

Column 2 introduces the incentive effect and shows an insignificant treatment effect on average. Specifically, the coefficient 0.09 translates into an incentive effect of 30 job entry points or approximately 5.6 people into employment per member of staff per year. Our parameter of interest is the interaction between treatment and office size (the number of staff). To control for the baseline productivity differences by the size of the office, Column 3 includes controls for a quadratic in mean frontline staff within an office, across time. Both terms are statistically significant and indicate that job entry productivity declines with staff at an increasing rate.

In an incentive scheme where performance is measured at a team level, the marginal return to individual effort decreases in the team size, which raises the incentive to free ride. Column 4 tests for the presence of free-riding within offices, including an interaction between treatment status and bins for office size. The omitted bin is for the middle half of the office size distribution, compared to those in the lower quartile (less than 12 frontline staff, bin 1) and those in the upper quartile (greater than 37 frontline staff, bin 3). Column 5 adds the number of offices per district. Column 6 controls for the differential district size (number of offices) in treated districts from the omission of PF offices by adding an interaction between treatment status and the number of offices per district to this specification.

Allowing for heterogeneity in the effect by the team size in columns 4-6 continue to show an insignificant mean effect. Relative to offices in the middle of team size distribution, small offices with few staff (in the bottom quartile) have significantly higher productivity but there is an insignificant incentive effect in large offices (in the top quartile). The coefficient for small offices translates into higher job entry productivity by 117 annual job entry points. Thus the average incentivisation effect of zero masks a positive mean effect in small offices. This effect is unchanged by controlling for both the number of offices per district (column 5) and falls only slightly when including the interaction between number of offices and incentivisation (column 6). The final row of Table 3 shows the p-value from a test for the equality of slope coefficients for the interaction of incentivisation with office size bins 1-3, indicating that we cannot reject significant heterogeneity in the incentive effect by team size at the 10% level. This supports the idea that free-riding is easier to mitigate in small offices by monitoring.[28][29]

3.2 Quality – Job-Seeker, Employer Service and Business Delivery

We adopt a similar approach to modelling quality outcomes. These are only measured quarterly and at district level. This reduces the sample size from around 900 offices to just 90 districts.[30] In terms of performance in the scheme, all incentivized districts met the quality target JSQ, 7 out of 17 districts met BDT target and 6 out of 17 met the EMQ target.

Columns 1-3 of Table 4 show the results for the district annual averages for JSQ, EMQ and BDT. Regional fixed effects are included in all regressions. Few variables are estimated to have a significant effect, due in part to the small number of observations and a lack of variation in the outcomes and in the case of JSQ the targets possibly having been set too low.

[28] We tested the robustness of these results to a change of the dependent variable to standardised *log* productivity to have mean 0 standard deviation 1 and standardised *level* of productivity. The size effects persist with the coefficient on incentivisation interacted with staff in bin 1 equal to 0.47 (0.49) and with staff in bin 3 equal to 0.26 (0.24) for standardised log (level) productivity. The corresponding standard errors are 0.23 (0.34) and 0.24 (0.14).

[29] We repeated the regression of Table 3 column 6 including PF offices to test for the sensitivity of our results to our choice to exclude these offices. The coefficient on the interaction term between treatment and staff in bin 1 and 3 changed to 0.27 and 0.09 respectively with standard errors 0.12 and 0.09, hence our conclusions did not change. In addition we ran a regression dropping Jobcentre Plus offices (offices that during the scheme became designated as pathfinder offices, predominantly in the final 2 months of the incentive scheme). The interaction between treatment and staff in bin 1 falls to 0.10 and is no longer statistically significant (standard error of 0.19) and the coefficient on the interaction of treatment and staff in bin 3 is very similar at 0.15 with standard error of 0.19. Only offices in incentivised districts were JCP offices and about half of the incentivsed offices were JCP, hence dropping these considerably reduces variation in the treatment status.

[30] We only received BDT data for 89 offices.

Similarly to above, an increase in the number of staff reduces productivity at an increasing rate. This may arise from a more personal service in smaller offices. For all quality measures the tightness of the labour market has a negative impact, the magnitude of which is largest for EMQ. This is intuitive, as a tight labour market means a difficult time for employers to fill vacancies. Importantly for our analysis, there is no significant impact of any term involving incentivisation status within small offices. There is a significant negative effect of incentivisation in large offices for JSQ, however the test of equality of slope coefficients cannot reject that the interactions of treatment with the three office size bins are equal to zero.

This lack of significant effect of incentivisation on quality outcomes can be taken in two different ways. On one hand, it could be argued that the scheme failed to elicit any increase due to design issues. The most obvious are the low precision of measurement and monitoring technology for quality. The theory says that low precision of measurement should lead workers to exert less extra effort on these tasks. Poor monitoring might make workers expect that any slack on effort would not be detected. Other explanations are that the targets were too low (but this is not supported by the fact that many districts did not hit the BDT and EMQ targets) or that the incentive payments were too low (but the fact that hitting all 5 targets represented between 7.5% to 8.5% of average pay and hitting 5 targets rather than 3 meant more than a doubling of the incentive payment does not support this). On the other hand, the failure of the scheme to have any effect on quality could be viewed more positively - as showing that despite the greater effort on quantity, quality did not actually fall, a standard failing of many incentive schemes. This may have been due to the fact that quality measures were explicitly part of the scheme and focusing only on the quantity target was not an option as the bonus payment was conditional on reaching at least two targets.

3.3 Quantity and Quality Together

We argue the contrast between the significant effect of the scheme on quantity and lack of effect arises partly from the differing measurement precision for quantity and quality. But it may simply be statistical as we have 90 observations (districts) in one case and over 800 (offices) in the other. We address this by re-running the quantity regression at district level using log district annual job entry productivity as the dependent variable. The results are in Column 4 of Table 4. There is a positive but insignificant impact of incentivisation on quantity, which increases in small offices. Again, we reject the hypothesis that the slope coefficients for the interaction between treatment and office size bins 1-3 are equal to zero at

the 10% level. This suggests that the differences between quantity and quality results are not due to size of the sample, but there is something different about the behavioural response to the quantity and quality targets.

It could be argued that since time allocated to quantity and quality is determined jointly, good performance on one will mean poor performance on the other. We therefore examine whether good performance on one dimension is positively or negatively correlated with good performance on the other. In fact, we find little correlation between quantity and quality, and a positive correlation between quality measures except EMQ and BDT[31]. If we take EMQ as more useful a measure (given the low variation in JSQ and BDT), there is a very low association[32]. Therefore we do not think that the results arise because time spent on quantity reduces the amount of time to achieve quality, but instead arise because of differences in measurement precision which means that aggregate output is much less related to individual productivity. These findings also have support from Gaynor and Pauly (1990) who showed that aggregate output was significantly higher in medical practices in which compensation was more closely related to individual productivity. In their context there is no joint production, as the organisation output is the number of office visits per week, observable at physician's level. So linking compensation to output is more effective. In our case we only have joint measures for output, and, for the quality measures, these are available only at district level, probably too weakly linked to individual productivity to be effective.

4. Valuing the impact of the incentive scheme

The mean effect of the scheme is zero. So across all offices in the scheme there were no gains. However, our estimates show that for small offices there were increases in quantity following the scheme. In this section, we ask the question if the scheme were to be introduced in the appropriate settings i.e. in small offices, what value might it have?

We can evaluate the quantitative importance of the change in the quantity outcome in two ways. First we compare the number of people placed into jobs with the monetary cost of the

[31] The correlation of average annual job entry productivity with JSQ, EMQ and BDT is 0.20, 0.13 and 0.24, between JSQ and EMQ and BDT is 0.38 and 0.49 and between EMQ and BDT is -0.041.

[32] We also estimate the district level annual quantity and EMQ models jointly using SUR, but as we would expect from the low correlation found above, there is only a small change in the standard errors.

scheme, thereby calculating the cost per placement. Second, we compare the benefits of the incentive scheme to the option of hiring more staff.

4.1 The cost per placement

Given the estimates in Table 3, column 6, we can straightforwardly calculate the distribution of change in job entry productivity associated with the incentive scheme. The fitted value from the regression is calculated using only variables related to treatment (the treatment effect itself plus any interactions), translated into job entry points then converted into a proportion of total job entry points for the treated office. Since the impact varies according to office size, we report this percentage change across the distribution as well as the mean in Table 5. As would be expected from Table 3 column 2, the overall effect of incentivisation is small at -0.169%. There is a substantial positive effect in small offices which declines across office size.

The mean percentage increase in small offices is 21%. A 21% change in job entry productivity translates to 19,979 job entry points. This is derived by calculating 21% of total job entry points relative to total staff for small, incentivized districts, multiplied by total staff in the small incentivized districts to give the total number of points. To convert points into people, we use Appendix Table W3 and normalise by the average points per job entry placement as 5.4 which gives 3,700 extra people. The *ex post* cost of the job entry component of the scheme was around £272,100. We estimate this from the data on the payments made for 5 of the 17 districts hitting their job entry target and earning 1% of salary (allowing for different numbers of staff). The figure is equivalent to 0.21% of the salary bill for the 17 incentivized districts. All 5 who hit were small districts. Consequently, in the best case scenario of targeting the incentive scheme towards small districts, the scheme cost £74 per job placement, a trivial amount.

4.2 Incentive Scheme versus Hiring More Staff

This can be compared with the option of hiring more staff. We ask how many staff would be needed to produce the additional 19,979 job entry points induced through the scheme. Mean productivity or job entry points per worker in the small, incentivized offices was 272.3 meaning that 73 extra staff members would be required to achieve the same effect of the incentive scheme. The average salary for EO and AO workers in 2002 was £15,515 and therefore the cost of hiring 73 more frontline staff would be £1,132,595 which is over 4 times

larger than the cost of the incentive scheme. The cost of the incentive scheme could hypothetically increase significantly from our estimate and still the incentive scheme would be considerably more cost effective than the option of hiring more staff.

5. Conclusions

There is little robust evidence on the role and impact of performance pay in the public sector, even though this is a sector that employs as many people in the UK and US as manufacturing does. This paper helps to start filling that gap by providing an evaluation of a pilot pay for performance scheme in a major UK government agency, Jobcentre Plus. The incentive scheme was based on team, rather than individual, performance and covered five different targets, measured with varying degrees of precision. We offer three main results: on the basic question of the efficacy of performance pay for public service workers; on the implications of the team basis of the scheme; and the implications of the explicit multiple targets including quality as well as quantity.

We show that the use of performance pay had no effect on average on the quantity measure (job placement productivity), but that there was important heterogeneity of response. The heterogeneity was patterned as one might expect from a free rider versus peer monitoring perspective. We found that the incentives had a substantial positive effect in small offices. In large offices, there was a negligible effect. Our interpretation of this is that peer monitoring and better information flows were able to overcome free rider problems in small units, but not in large teams.

The impact of performance pay on quantity was not matched by any impact on quality measures. Our interpretation of this finding is that individuals responded optimally to the scheme by focusing their effort on the better-measured quantity outcome rather than quality. It seems likely that the main aim of the incentive scheme was to raise quantity, and the introduction of quality targets was to mitigate or prevent declines in quality rather than in the expectation of improvements in quality and this proved successful.

There are, of course, a number of caveats we need to note. First, the scheme only operated for one year and so the responses may include a “first year” novelty effect in addition to the pure

incentive effect. Furthermore, if a 'ratchet' design of continual percentage improvements were repeated in a dynamic setting, the optimal response would be different to the response from a once-only pilot. Second, the outcome could be the result of performance management per se, rather than the financial reward attached. However, this is unlikely. The same performance management system was in place everywhere, in both control and treated offices. It may be that the financial incentives led managers to take the existing framework more seriously but that is surely part of the aim of performance pay. Third, given the specific structure of the agency and the incentive scheme, it is possible to question the external validity of this study, but there is little evidence on the use of incentive schemes in the public sector with which to compare our findings.[33] While there is more evidence for the private sector, the differences between the sectors mean that schemes in the two sectors are sufficiently different to limit comparison.[34] Fourth, our identification strategy removes Pathfinder offices from the evaluation sample, These offices were not selected on a productivity basis, but rather given their ability to cope well with the new structure of Jobcentre Plus. If however the selection criteria was correlated with productivity, then dropping the most productive offices from incentivized districts may produce a lower bound on the treatment effect. We think that the quantitative impact would be second order at best, as there were a substantial number of offices in each district so removing one could only have a trivial effect.[35]

Finally, we draw some tentative conclusions for the design of team-based performance pay schemes in the public sector. These relate to both incentive scheme design and organisational design. One key lesson is that designing a sensible performance pay system in the public sector is difficult. A common problem is the strongly hierarchical structure often present and the resulting difficulty of attributing outcomes across different levels. One way forward

[33] One notable exception is the experimentation with various forms of performance related pay in the JTPA schemes in the USA but these did not have the same design features which allowed examination of free-riding. For example Heckman et al. 1996 examined the prevalence of cream-skimming and Courty and Marschke, 1997 found bureaucrats to manipulate the date of reporting outcomes of participants to maximise own bonus payments.

[34] For example, Bandiera et al. 2005, 2007, 2009 run a field experiment in a private firm and consider the impact of relative incentives versus individual incentives on workers' productivity and on managers incentives to select workers. In the public organization we consider, the scheme designers wanted to promote cooperation across workers, so that the use of relative performance evaluation, which puts workers in competition to each other, would have not been considered. It is also more likely that in a public organisation there is more pressure from unions to use group performance evaluation rather than individual performance evaluation.

[35] If we exclude the largest office from non-incentivized districts for example, mean job entry productivity only changes by 3.7%. This compares to a standard deviation in office job entry productivity which is 170% of the mean.

would be to explicitly recognise that different levels of performance are all valued and to offer incentives as a weighted function of individual-, group-, and organization-level performance as is common in the private sector[36]. Another option of course is to measure all outputs at the most disaggregated level although this may be a costly solution. We have shown that the scheme was very cost-effective in small offices, despite not being particularly high powered (in contrast, for example, to Kahn et al. 2001). Ideally then team size needs to be small and preferably not dispersed over many sites, and the connection between effort and output needs to be as clear and well-measured as possible. The argument of Dewatripont, Jewitt and Tirole (1999) about organisational design around missions can be adapted here. If incentives are indeed a cost-effective way of inducing greater output given the right team size, then it may make sense to re-structure organisations to create natural teams of the appropriate size. Such re-structuring could also allow for relative performance evaluation to filter away common uncertainty and fits well with the general movement towards devolved agency inherent in many current public service reforms.

[36] We thank a referee for this suggestion.

References

Abbink, K., Irlenbusch, B. and Renner, E., 2006. Group size and social ties in microfinance institutions, Economic Inquiry, Western Economic Association, 44(4), 614-628.

Baiker, K., Jacobson, M., 2007. Finders keepers: forfeiture laws, policing incentives, and local budgets. Journal of Public Economics 91, 2113-2136.

Baker, G., 2002. Distortion and risk in optimal incentive contracts. Journal of Human Resources 37(3), 728-751.

Bandiera, O., Barankay, I. and Rasul, I., 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. The Quarterly Journal of Economics, 120(3), 917-962.

Bandiera, O., Barankay, I. and Rasul, I, 2007. Incentives for Managers and Inequality Among Workers: Evidence From a Firm-Level Experiment. The Quarterly Journal of Economics, 122(2), 729-773.

Bandiera, O., Barankay, I. and Rasul, I, 2009. Social Connections and Incentives in the Workplace: Evidence From Personnel Data. Econometrica, 77(4), 1047-1094.

Bandiera, O., Guiso, L., Prat, A., Sadun, R., 2011. Matching firms, managers and incentives. NBER Working Paper 16691.

Barnow, B. A., 2000. Exploring the relationship between performance management and program impact: A case study of the JTPA. Journal of Policy Analysis and Management 19(1), 118-141.

Besley, T., Ghatak, M., 2003. Incentives, choice and accountability in the provision of public services. Oxford Review of Economic Policy 19(2), 235 – 249.

Besley, T., Ghatak, M., 2005. Competition and incentives with motivated agents. American Economic Review 95(3), 616-636.

Bhattacherjee, D., 2005. The effects of group incentives in an Indian firm: evidence from payroll data. Review of Labour Economics and Industrial Relations 19(1), 147-173.

Bloom, N., Van Reenen, J., 2010. Human resource management and productivity. NBER Working Paper 16019 and forthcoming. In: Ashenfelter, O and Card, D. Handbook of Labor Economics 4.

Burgess, S., Propper, C., Ratto, M.L., von Hinke Kessler Scholder, S., Tominey, E., 2010. Smarter task assignment or greater effort: what makes a difference in team performance? The Economic Journal 120(547), 968-989.

Burgess, S., Ratto, M.L., 2003. The role of incentives in the public sector: issues and evidence. Oxford Review of Economic Policy 19(2).

Carpenter, J., 2007. Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. Games and Economic Behavior 60(1), 31-51.

Courty, P and Marschke, G., 1997. Measuring government performance: lessons from a federal job training programme. American Economic Review 87(2), 383-388.

Dewatripont, M., Jewitt, I., Tirole, J., 1999. The economics of career concerns, Part II: Application to missions and accountability of government agencies. The Review of Economic Studies 66(1), 199-217.

Dixit, A., 2002. Incentives and organisations in the public sector: an interpretative review. Journal of Human Resources 37(4), 696-727.

Francois, P., 2000. Public service motivation' as an argument for government provision. Journal of Public Economics 78, 275-299.

Gaynor, M., Pauly, M., 1990. Compensation and productive efficiency in partnerships: evidence from medical group practice. Journal of Political Economy 98(3), 544-573.

Gaynor, M., Rebitzer, J.B., Taylor, L.J., 2004. Physician incentives in health maintenance organizations. Journal of Political Economy 112(4), 915-931.

Gertler, P. and Vermeersch, C. 2012. Using performance incentives to improve health outcomes. Policy Research Working Paper Series 6100, The World Bank.
Ghatak, M. and Guinnane, T., 1999. The economics of lending with joint liability: theory and practice. Journal of Development Economics, 60, 195-228.
Gravelle, H., Sutton, M., Ma, A., 2010. Doctor behaviour under a pay for performance contract: treating, cheating and case finding? The Economic Journal 120, 129-156.
Hamilton, B.H., Nickerson, J.A., Owan, H., 2003. Team incentives and worker heterogeneity: an empirical analysis of the impact of teams on productivity and participation. Journal of Political Economy 111(3), 465-497.
Heckman, J., Smith, J., Taber, C., 1996. What do bureaucrats do? The effects of performance standards and bureaucratic preferences on acceptance into the JTPA program. In: G. Libecap. (Eds.). Advances in the study of entrepreneurship, innovation and growth. 7. Greenwich, CT: HAI Press, 191-217.
Holmström, B., 1982. Moral hazard in teams. Bell Journal of Economics 13, 324-340.
Holmström, B., Milgrom, P., 1991. Multi-task Principal-Agent problems: incentive contracts, asset ownership and job design. Journal of Law, Economics and Organization, 7 (Special Issue), 24-52.
Kahn, C.M., Silva, E.C.D., Ziliak, J.P., 2001. Performance-based wages in tax collection: The Brazilian tax collection reform and its effects. Economic Journal 111(468), 188-205.
Kandel, E., Lazear, E., 1992. Peer pressure and partnerships. Journal of Political Economy 100(4), 801-817.
Knez, M., Simester, D., 2001. Firm-wide incentives and mutual monitoring at Continental Airlines. Journal of Labor Economics 19(4), 743-772.
Lavy, V., 2009. Performance pay and teachers' effort, productivity, and grading ethics. American Economic Review 99(5), 1979-2011.
Lazear, E., 2001. Performance pay and productivity. American Economic Review 90(5), 1346-1361.
Makinson, J., 2000. Incentives for change. Rewarding performance in national government networks. Public Service Productivity Panel. HMSO.
Mullen, K., Frank, R., Rosenthal, M., 2010. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. The RAND Journal of Economics 41(1), 64-91.
Muralidharan, K. and Sundararaman, V., 2011. "Teacher performance pay: experimental evidence from India," Journal of Political Economy, 119(1), 39 - 77
Nickell, S., Quintini, G., 2002. The consequences of the decline in public sector pay in Britain: a little bit of evidence. Economic Journal 112(477), 107-118.
Osborne, D., Gaebler, T., 1993. Reinventing government: how the entrepreneurial spirit is transforming the public sector. New York: Plume Books, (Penguin Group).
Paarsch, H., Shearer, B., 2000. Piece rates, fixed wages, and incentive effects: statistical evidence from payroll records. International Economic Review 41(1), 59-92.
Prendergast, C., 1999. The provision of incentives in firms. Journal of Economic Literature 37, 7-63.
Prendergast, C., 2002. The tenuous trade-off between risk and incentives. Journal of Political Economy 110(5), 1071-1102.
Propper, C., Van Reenen, J., 2010. Can pay regulation kill? Journal of Political Economy 118(2), 222-273
Ratto, M., Tominey, E., Vergé, T., 2010. Rewarding collective performance to induce cooperation. Mimeo University of Bristol.
White Paper, 1999. Modernising Government. www.archive.official-documents.co.uk
World Development Report, 2003. http://econ.worldbank.org/wdr/wdr2003/

Figure 1: Distribution of Annual Log Job Entry Productivity

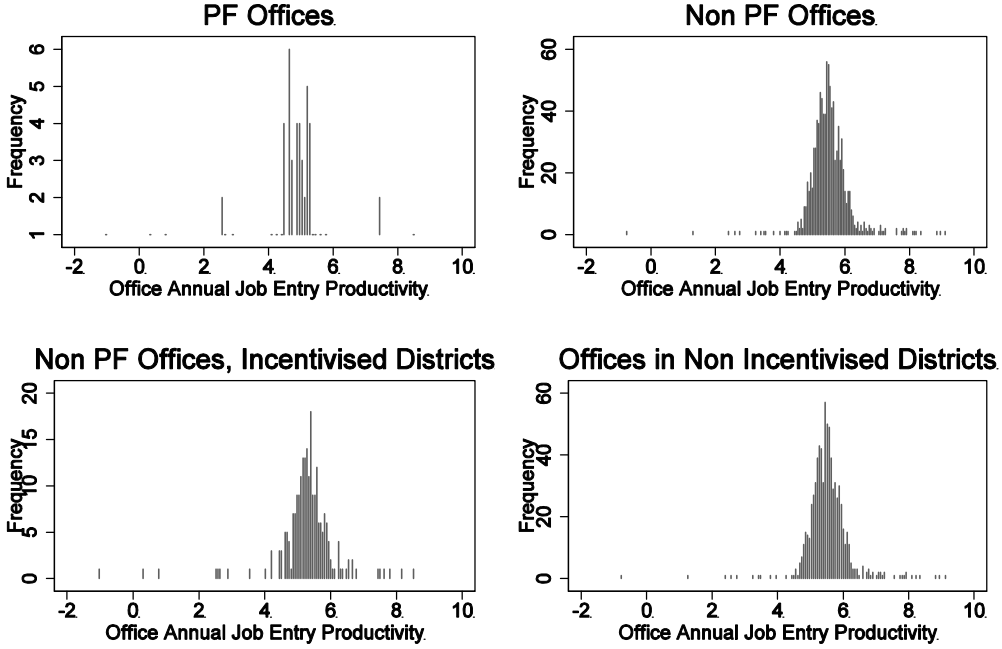

Figure 2: Distribution of District Quality Measures

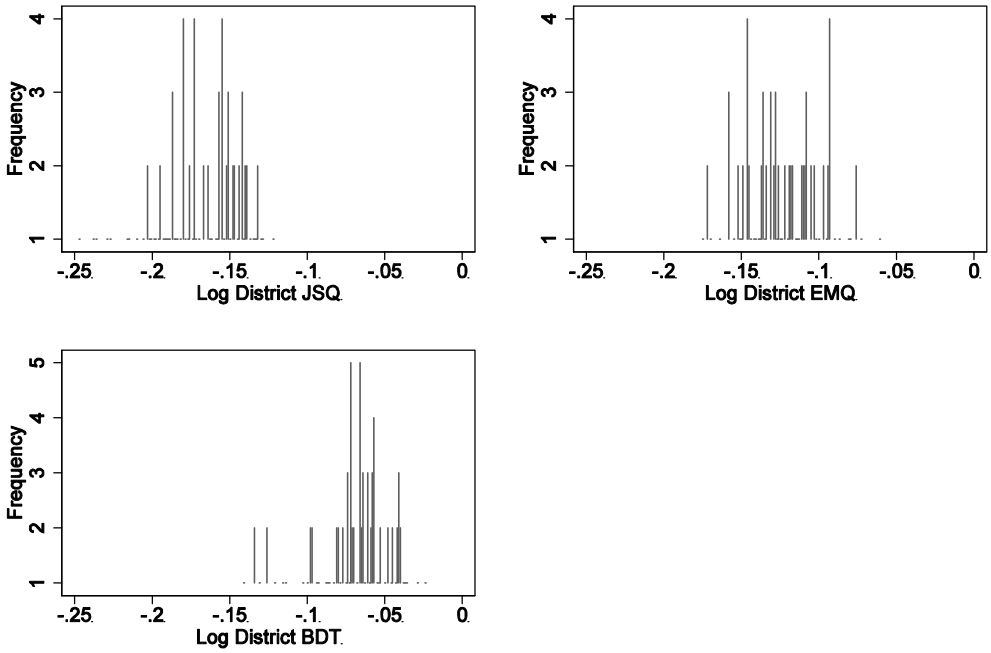

Table 1: Characteristics of the districts and offices by incentive status

(a) Districts

		% Pathfinder Office	Frontline Staff	Mean Labour Market Conditions	Private Public Wage Gap	Number of Offices in District
Non-Incentivized Districts	Mean		268.38	1.29	-0.27	11.14
	Median		235	1.23	-0.85	11
Incentivized Districts	Mean	11.84	424.82	1.18	-0.21	14.54
	Median	11.77	405	1.14	-0.77	13.25
Incentivized Districts, excluding PF offices	Mean	11.84	242.08	1.20	-0.35	10.84
	Median	11.77	204	1.21	-0.66	9

(b) Offices

		Pathfinder Office	Frontline Staff	Mean Labour Market Conditions	Private Public Wage Gap
Offices in Non-Incentivized Districts	Mean		27.02	1.21	-0.21
	Median		21	1.11	-0.77
Offices in Incentivized Districts	Mean	0.22	32.01	1.18	-0.17
	Median	0	24	1.13	-0.59
Non PF Offices in Incentivized Districts	Mean		25.16	1.19	-0.17
	Median		21	1.15	-0.68

Note: Frontline staff defined as the sum of Executive Officer and Administrative Officer. Labour market is defined as the ratio of the inflow of unemployment claimants to the inflow of vacancies, by TTWA.

Table 2: Balancing tests for Propensity Score Match Quality

		Mean			% reduction	t-test	
Variable	Sample	Treated	Control	%bias	|bias|	t	p>|t|
Propensity Score	Unmatched	0.24	0.13	63.70		7.38	0.00
	Matched	0.24	0.23	3.30	94.90	0.19	0.85
Log Mean Frontline Staff	Unmatched	2.93	3.01	-9.90		-0.91	0.36
	Matched	2.93	3.07	-16.80	-70.20	-1.17	0.24
Office Frontline Staff Variance	Unmatched	3.09	2.67	9.80		0.85	0.40
	Matched	3.09	3.03	1.30	86.90	0.09	0.93
Office Frontline Staff Squared	Unmatched	931.92	1455.20	-12.40		-0.90	0.37
	Matched	931.92	1184.40	-6.00	51.70	-0.90	0.37
Frontline Staff * No. Offices	Unmatched	332.33	336.80	-1.70		-0.15	0.88
	Matched	332.33	399.06	-24.80	-1393.40	-1.71	0.09
Office Log Mean Labour Market	Unmatched	0.08	0.13	-16.90		-1.46	0.14
	Matched	0.08	0.11	-8.40	50.50	-0.67	0.51
Office Labour Market Variance	Unmatched	0.29	0.33	-23.70		-1.93	0.05
	Matched	0.29	0.30	-5.70	76.10	-0.49	0.62
Labour Market * Frontline Staff	Unmatched	27.55	31.81	-17.30		-1.48	0.14
	Matched	27.55	31.35	-15.50	10.70	-1.28	0.20
No. Offices per District	Unmatched	13.44	12.67	19.40		1.74	0.08
	Matched	13.44	14.06	-15.50	20.20	-1.07	0.29
No. Offices per District Squared	Unmatched	194.16	178.58	14.10		1.29	0.20
	Matched	194.16	217.17	-20.80	-47.60	-1.40	0.16
Proportion of High Grade Staff	Unmatched	0.03	0.03	-2.10		-0.20	0.84
	Matched	0.03	0.03	-1.10	46.60	-0.08	0.93
Private Public Wage Gap	Unmatched	0.24	-0.13	13.90		1.42	0.16
	Matched	0.24	-0.47	26.90	-93.80	1.85	0.07

Note: Offices included in analysis contributed towards job entry outcome, but were non Pathfinder offices. Mean labour market is defined as the monthly ratio of the inflow of unemployment claimants to the inflow of vacancies, by TTWA, averaged within offices across time. Mean frontline staff defined as the sum of Executive Officer and Administrative Officer, averaged within offices across time. The t-test and p-value are reported for the test of equality of means in treated and control observations.

Table 3: Office Annual Job Entry Productivity

	(1)	(2)	(3)	(4)	(5)	(6)
District Office	0.03	0.03	0.06	0.07	0.07	0.06
	(0.063)	(0.087)	(0.080)	(0.079)	(0.079)	(0.080)
JCP Office	0.04	-0.04	0.06	0.07	0.07	0.05
	(0.098)	(0.128)	(0.117)	(0.118)	(0.119)	(0.120)
Private Public Wage Gap	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	(0.017)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)
Log Mean Labour Market	0.09	0.09	-0.03	-0.04	-0.04	-0.04
	(0.061)	(0.081)	(0.074)	(0.074)	(0.074)	(0.074)
Incentivisation Status		0.09	-0.01	-0.17	-0.17	0.16
		(0.095)	(0.087)	(0.106)	(0.107)	(0.292)
Mean Office Frontline Staff/100			-1.50***	-1.49***	-1.49***	-1.48***
			(0.134)	(0.138)	(0.140)	(0.140)
Mean Office Frontline Staff/100 Squared			0.27***	0.26***	0.26***	0.26***
			(0.058)	(0.059)	(0.059)	(0.059)
Incentivisation * Mean Frontline Staff Bin 1 (12 staff)				0.35**	0.35**	0.30**
				(0.140)	(0.142)	(0.147)
Incentivisation * Mean Frontline Staff Bin 3 (37 staff)				0.21	0.21	0.17
				(0.149)	(0.149)	(0.153)
No. Offices Per District					0.00	0.00
					(0.006)	(0.006)
Incentivisation * No. Offices Per District						-0.02
						(0.018)
Constant	5.48***	5.48***	5.86***	5.86***	5.84***	5.82***
	(0.032)	(0.093)	(0.091)	(0.091)	(0.112)	(0.113)
Observations	841	841	841	841	841	841
R-squared	0.077	0.078	0.234	0.240	0.240	0.241
P-value for Test Equality Slope Coefficients				0.089	0.086	0.060

Note: standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Dependent variable: Log Office Productivity. Office Productivity is defined as log of the ratio of total JE points to total frontline staff. Pathfinder offices were omitted from analysis. All regressions control for regional fixed effects. The district office indicates the head office. JCP status is a dummy variable equal to one if the offices within incentivized districts were given the new JCP status during the incentive scheme and 0 otherwise. The private public wage gap is defined as the relative hourly wage differential within Local Authorities. Mean labour market is defined as the monthly ratio of the inflow of unemployment claimants to the inflow of vacancies, by TTWA, averaged within offices across time. Mean frontline staff defined as the sum of Executive Officer and Administrative Officer, averaged within offices across time. Bin 1 of staff is less than 12, bin 3 is greater than 37 (25^{th} and 75^{th} percentile respectively). Regressions are clustered at the district level.

Table 4: District annual JSQ, EMQ, BDT and JE analysis

	(1) JSQ	(2) EMQ	(3) BDT	(4) JE
% PF offices per District	0.00	-0.01	-0.01	-0.09
	(0.005)	(0.007)	(0.005)	(0.088)
% JCP offices per District	-0.00	-0.00	-0.00	0.01
	(0.003)	(0.004)	(0.002)	(0.029)
Private Public Wage Gap	-0.00	-0.00	0.00	-0.02
	(0.002)	(0.003)	(0.002)	(0.026)
Mean District Frontline Staff/100	-0.12*	-0.19**	0.05	-1.74**
	(0.064)	(0.078)	(0.058)	(0.756)
Mean District Frontline Staff/100 Squared	0.08	0.16**	-0.03	0.74
	(0.055)	(0.067)	(0.045)	(0.616)
Log Mean District Labour Market	-0.03*	-0.06***	-0.01	-0.10
	(0.016)	(0.016)	(0.011)	(0.157)
Incentivisation Status	-0.02	0.11	0.09	0.70
	(0.054)	(0.080)	(0.063)	(1.099)
Incentivisation * Mean District Frontline Staff Bin 1 (12 staff)	0.00	-0.00	-0.00	0.30**
	(0.007)	(0.011)	(0.007)	(0.141)
Incentivisation * Mean District Frontline Staff Bin 3 (37 staff)	-0.02**	0.01	-0.01	0.09
	(0.009)	(0.013)	(0.011)	(0.130)
No. Offices per District	0.00	-0.00*	0.00	-0.01
	(0.001)	(0.001)	(0.001)	(0.006)
Incentivisation * No. Offices per District	-0.00	-0.00	-0.00	0.01
	(0.001)	(0.001)	(0.001)	(0.013)
Constant	-0.12***	-0.01	-0.08**	3.46***
	(0.044)	(0.052)	(0.036)	(0.457)
Observations	90	90	89	90
R-squared	0.663	0.435	0.587	0.688
P-value for Test Equality Slope Coefficients	0.167	0.446	0.109	0.086

Note: Standard errors in parentheses. * significant at 10%; ** significant at 5%; ** significant at 1%. JSQ is the job seekers service, EMQ the employer quality outcome, BDT business delivery target and JE job entries. Dependent variables are log annual district average JSQ, log annual district average EMQ, log annual district average BDT and log productivity of JE. All regressions include regional fixed effects. PF denotes the Pathfinder Office created prior to the incentive scheme. JCP status is a dummy variable which equals one if the offices within incentivized districts were given the new JCP status during the incentive scheme and 0 otherwise. The private public wage gap is defined as the relative hourly wage differential within Local Authorities. Mean frontline staff defined as the sum of Executive Officer and Administrative Officer, averaged within offices across time. Mean labour market is defined as the monthly ratio of the inflow of unemployment claimants to the inflow of vacancies, by TTWA, averaged within offices across time. Bin 1 of staff is less than 12, bin 3 is greater than 37 (25^{th} and 75^{th} percentile respectively). Regressions are clustered at the district level.

Table 5: Mean Incentivisation Effect

Number of frontline staff per office, by bin for office size			
<=12 staff	12-37 staff	>=37 staff	Mean
21.431	-15.219	5.361	-0.169

Note: The incentivisation effect for incentivized offices was calculated using the fitted value from Table 3, column 5, using the variables incentivisation status and an interaction of this with office size and district size.

Appendix

Table A1: Data Descriptives

Variable	**Mean**	**Standard Deviation**		
		Total	**Between**	**Within**
Office Level Variables				
Office Monthly Job Entry Points	512.893	458.300	458.300	183.384
Office Pathfinder Status	0.054	0.226	0.224	0.000
Office JCP Status	0.062	0.241	0.256	0.030
District Office	0.056	0.229	0.244	0.000
Private Public Wage Gap	-0.550	2.417	2.382	0.000
Log Office Frontline Staff	3.085	0.856	0.857	0.197
Office Frontline Staff Variance	4.503	6.394	6.163	0.000
Log Office Labour Market	0.091	0.441	0.311	0.312
Incentivisation Status	0.241	0.428	0.426	0.000
Office Mean % High Grade Staff	0.034	0.034	0.034	0.000
Labour Market Time Series Variation	0.262	0.057	0.057	0.000
District Level Variables				
Log District Annual Job Entry Points	13.818	0.359	0.361	0.000
Log District EMQ	-0.146	0.044	0.027	0.034
Log District JSQ	-0.170	0.035	0.028	0.022
Log District BDT	-0.074	0.026	0.025	0.009
% PF Offices per District	2.535	4.941	4.659	1.131
% JCP Offices per District	0.005	0.022	0.016	0.017
Log District Mean Frontline Staff	8.357	0.462	0.451	0.000
Log District Labour Market	0.130	0.353	0.217	0.285
No. Offices per District	11.647	3.981	4.265	0.000
M * No. Offices per District	3.456	6.608	6.603	0.000

Table A2: Propensity score probit estimates of Incentivisation Status

Mean Frontline Staff	-0.265*
	(0.149)
Office Frontline Staff Variance	-0.004
	(0.014)
Office Frontline Staff Squared	-0.000*
	(0.000)
Frontline Staff * No. Offices	0.001
	(0.001)
Office Mean Labour Market	0.502
	(0.394)
Office Labour Market Variance	-0.706
	(0.541)
Labour Market * Frontline Staff	0.008
	(0.009)
No. Offices per District	-0.660***
	(0.086)
No. Offices per District Squared	0.028***
	(0.003)
Proportion of High Grade Staff	0.605
	(2.134)
Private Public Wage Gap	0.014
	(0.039)
Regional Variables	
East of England	0.020
	(0.369)
London	0.913**
	(0.407)
North East	0.243
	(0.468)
North West	0.817**
	(0.345)
Office for Scotland	0.842**
	(0.337)
Office for Wales	0.482
	(0.361)
South East	-3.608***
	(0.529)
South West	-0.481
	(0.395)
West Midlands	0.681*
	(0.353)
Yorkshire	-0.124
	(0.408)
Constant	2.394***
	(0.751)
Observations	912
Psuedo R^2 = 0.2813	

Notes: Standard errors in parentheses * significant at 10%; ** significant at 5%; *** significant at 1%. Note: the predicted value forms the propensity score used for the office level quantity analysis. Offices included in analysis contributed towards job entry outcome, but were non Pathfinder offices. Labour market is defined as the ratio of the inflow of unemployment claimants to the inflow of vacancies, by TTWA. Mean frontline staff defined as the sum of Executive Officer and Administrative Officer, averaged within offices across time.

Table A3: Propensity Matching Results

Sample on support	Average Treatment of the Treated
841	0.002
	(0.085)

Note: Outcome is log job entry productivity per office. This is calculated as the log of total office job entry points per staff. Pathfinder offices are omitted from analysis. Kernel weighted propensity score matching with an Epanechnikov kernel. Bootstrapped standard error in parentheses, with 100 replications